Transforming Object Locations on a 2D Visual Display Into Cued Locations in 3D Auditory Space

Anthony Hornof, Tim Halverson, Andy Isaacson, and Erik Brown
University of Oregon
Eugene, Oregon, USA

An empirical study explored the extent to which people can map locations in auditory space to locations on a visual display for four different transformations (or mappings) between auditory and visual surfaces. Participants were trained in each of four transformations: horizontal square, horizontal arc, vertical square, and vertical spherical surface. On each experimental trial, a sound was played through headphones connected to a spatialized sound system that uses a non-individualized head-related transfer function. The participant's task was to determine, using one transformation at a time, which of two objects on a visual display corresponded to the location of the sound. Though the two vertical transformations provided a more direct stimulus-response compatibility with the visual display, the two horizontal transformations made better use of the human auditory system's ability to localize sound, and resulted in better performance. Eye movements were analyzed, and it was found that the horizontal arc transformation provided the best auditory cue for moving the eyes to the correct visual target location with a single saccade.

Auditory displays are routinely used to keep an operator abreast of what is happening in the visual periphery. Auditory alerts often direct attention to visual displays in cars, aircraft, and computer interfaces. Though characteristics of sound such as pitch, timbre, and timing are good for conveying specific encodings (Gaver, 1997), the physical location of an auditory alert in three-dimensional space can also convey useful meaning. The location of an auditory alert in three-dimensional (3D) space could, for example, help an air traffic controller to direct his or her visual attention to a particular blip on a radar screen.

Previous research has examined the extent to which people can discriminate the precise location of auditory stimuli. There are a range of results in terms of how accurately people can locate a sound in space, depending on a range of experimental conditions, such physical versus virtual localization (Wightman & Kistler, 1989), the use of non-individualized head-related transfer functions (Wenzel, Arruda, Kistler, & Wightman, 1993), and egocentric versus exocentric localization (Simpson, et al., 2007). In general, people can distinguish the locations of auditory sound sources better when there is greater separation between them, requiring roughly 9° of azimuth or 12° of elevation (Begault, 1994, p. 67; Grantham, Hornsby, & Erpenbeck, 2003).

Additional research has investigated the utility of spatialized audio to locate visual targets. In general, people can distinguish the locations of aurally-cued visual targets better when the visual display is sparse and the auditory cues are reliable (Perrott, Sadralodabai, & Saberi, 1991; Vu & Strybel, 2006). However, little if any research has explored the potential benefits of transforming a small visual region into a larger auditory space. If spatialized audio is to be used to direct visual attention to a location on a small visual display, the best spatial resolution might be obtained if the visual display is expanded and transformed into a larger auditory space, but there are many possible ways to make this transformation.

The experiment presented here explores the extent to which people can map locations in auditory space to locations in visual space for four different transformations (or mappings) between auditory and visual space. The goal is to provide a specific recommendation for how to best convey the location of an object on a 2D visual display using 3D audio.

METHOD

The primary task consisted of a forced-choice discrimination task between two visual objects, or "blips," that appeared on a simulated radar on the left side of the visual display. This was interleaved with a secondary tracking task that consisted of using a joystick to keep a set of crosshairs on a target that moved around on the right side of the visual display. Just before the appearance of the two blips on the left side of the display, a sound was played with its spatial location indicating which of the two blips was the target. The participant quickly keyed in the number located on the target blip. Four different visual-to-auditory transformations were used, one at a time.

Participants

Sixteen graduate and undergraduate students participated. The mean age was 23 years. Seven were women. All had normal or corrected-to-normal vision, no known hearing deficiencies, and use of both hands. All considered themselves to be right-handed. Participants were paid $12 plus up to $7.40 in bonuses, to motivate speed and accuracy. Each experimental session lasted roughly one hour.

Apparatus

Visual Stimuli. Visual stimuli appeared in two distinct regions on a computer display positioned 61 cm from the participant. The primary visual region was a rectangle that subtended 16° (horizontal) by 13° (vertical) visual angle, just left of center on the display. The blips in this region were yellow bullet-shaped blips from Smallman et al. (2001), alongside each of which appeared a *1*, *2* or *3*. The secondary visual region was a square that subtended 14° visual angle horizontally and vertically, just right of center on the computer display. The object to track in the secondary region was a small red circle. Stimuli were presented (and responses recorded) using an Apple Macintosh G5 dual processor system and experimen-

tal software written in the C++ programming language using Apple XCode.

For the blip identification task in the primary visual region, forty pairs of blip locations were randomly generated, such that the two blips were always at least 1° of visual angle from the edge of the region and at least 4° of visual angle from each other. The same forty pairs of locations were used for each of the four visual-to-auditory mappings. Pair-presentation orderings were generated randomly, and the "correct" blip of each pair was randomly assigned in each ordering.

Spatialized Audio. Spatialized audio was generated using a VRSonic SoundSim Cube spatialized audio server and Sennheiser HD250-II headphones. The head-related transfer function (HRTF) used to spatialize sounds for all participants was CIPIC HRTF #158, which was chosen based on a preliminary evaluation in which the four authors blindly selected this HRTF as providing the best spatialization that is "out of the head" and with clear front-back distinction.

Visual to Auditory Transformations. Figure 1 shows the four visual-to-auditory transformations, or mappings, used in the experiment. The two horizontal transformations primarily use azimuth and intensity to convey the object location. The two vertical transformations primarily use azimuth and elevation. The two squares map the Cartesian grid of the visual display directly onto a grid in the audio space, resulting in a slight covariation of intensity with other dimensions. The two curved displays map the horizontal component of the visual display to azimuth, and the vertical component to either distance or elevation.

Each auditory plane or surface is defined with specific measurements that define where the sounds are placed in virtual space. An effort was made to normalize the four auditory displays such that, for example, objects appeared at roughly the same distances for each transformation. However, a more important consideration was to select dimensions that best conveyed location information. For example, the vertical spherical surface needs to be high and wide to maximize localization, but cannot be a complete half-sphere because in this case all items near the top would converge at the same point. We arrived at the following dimensions: The horizontal square ranges from 5 to 20 m in front of the listener. The horizontal arc subtends 180° at a distance from 5 to 20 m. The vertical square is 8 m in front of the listener and subtends 140° on the horizon. The vertical spherical surface is at a distance of 10 m and subtends 120° for both elevation and azimuth.

Alert Sound. The alert sound was P2 from Cabrera, Ferguson, and Laing (2005), shortened to three repetitions of the tone (duration = 1.55 s). The sound was akin to an electronic bell or "dong" sound. Figure 2 shows a frequency analysis of the sound (from Audacity 1.2.6). Each successful trial was rewarded with a pleasant "cha-ching" cash register sound, whereas each error evoked an annoying buzzer, both 5 m directly in front of the listener.

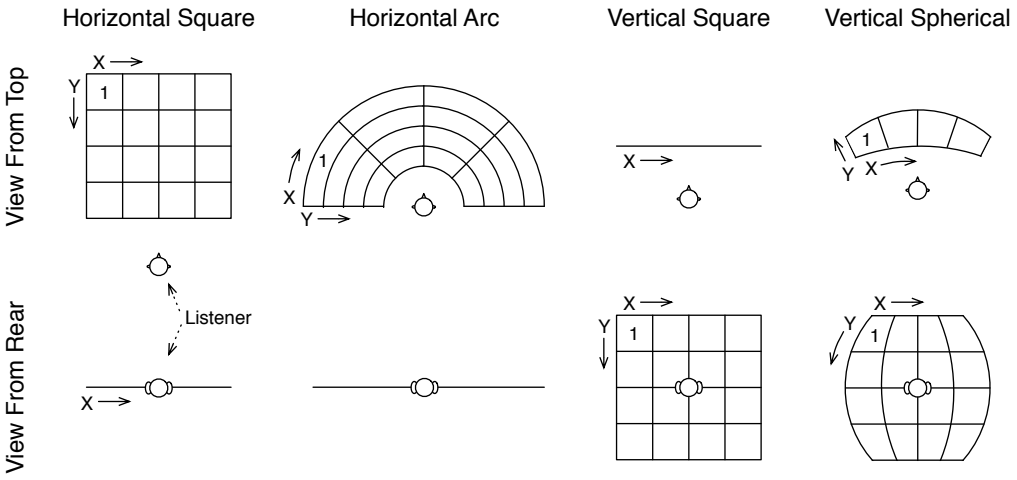

Figure 1. The visual display was transformed to these four auditory surfaces. The diagrams show how each auditory surface was mapped to a visual display with a horizontal X axis, vertical Y axis, and a hypothetical object 1 in the top right corner.

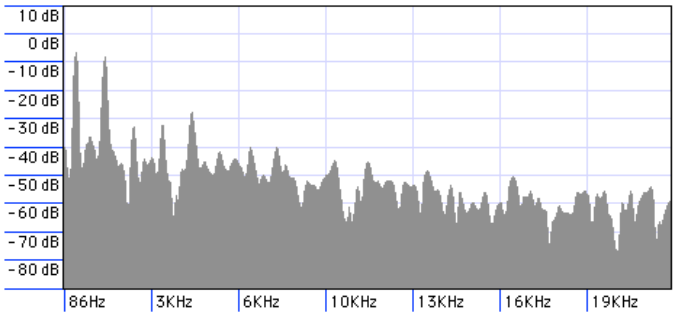

Figure 2. A frequency analysis of the alert sound used in the study. The second major peak is at 1,570 Hz.

Eye Tracking. We tracked each participant's gaze so that we could report, after a sound is played, (a) how long it took for participants to move their gaze to the primary visual display and (b) to what location on the visual display participants made their first eye movement (as in, how close to the visual location that corresponds to the sounded auditory location). Gaze was tracked using an LC Technologies 120 Hz binocular pupil-center corneal-reflection Eyegaze eye tracker, with an accuracy of roughly 20 ms and 1° of visual angle. A chinrest was used to keep the eyes in range of the eye tracking cameras. Eye movement measures were analyzed using the VizFix eye movement analysis tool developed in our lab. Fixations were determined with a dispersion-based algorithm, assuming a 100 ms minimum fixation duration and a maximum gaze sample dispersion radius of 0.5° of visual angle.

Procedure

After eye tracker calibration, the participant was presented with four visual-to-auditory transformations. The order of transformations was randomized across participants using a Latin square. For each transformation, the mapping was explained both verbally and physically with objects that showed the surfaces in auditory space. The participant then watched a sequence of blips on the visual display and listened to the corresponding locations in the auditory display. The participant

Figure 3. The experimental setup including the visual display, chinrest, keypad, headphones, joystick, and a physical representation of the horizontal arc transformation.

could review this sequence as many times as desired. Figure 3 shows the physical setup, including a board cut out to show the horizontal arc transformation around a miniature "listener."

Following each training session, the timed portion of the trial commenced. The participant started the tracking task in the secondary visual region, in which a blip wandered around the screen and the participant was instructed to keep a crosshair on the blip using a joystick. The participant was told to return to the tracking task as much as possible and that good tracking accuracy would earn them an additional two dollars at the end of the experiment.

When a spatialized alert sounded, the participant moved their gaze to the primary visual region on the left side of the screen, decided which of the two blips had sounded, and entered that blip number with their left hand on a numeric keypad. The participant had three seconds to press the *1*, *2* or *3* key (typically with their middle or first finger) and *Enter* (typically with their thumb). If a correct identification was entered within three seconds, the reward sounded and the participant earned three cents (as they were told in advance). If an incorrect identification was entered, the penalty sounded. Blips remained on screen for three seconds, after which three to seven seconds elapsed before the presentation of the next two blips and auditory alert. During these intervals, the participant resumed the tracking task.

Both the primary and secondary displays were *gaze-contingent.* That is, visual objects appeared in each of the two task windows only when the eyes were in that window. The blips appeared roughly 20 ms after the gaze arrived in each window, which was barely perceptible and not at all distracting. Since participants were instructed and financially motivated to keep their gaze on the tracking task in the secondary display, the gaze-contingent design meant that participants perceived the auditory alerts *before* the visual stimuli.

This procedure was repeated for each of the four transformations with an optional break between each block, after which participants were interviewed regarding their subjective impressions. The primary measures of interest were, for each transformation, how quickly and accurately participants could determine which of the two blips the sound corresponded to, and which transformation provided the best assistance in moving the eyes to the target location with a single saccade. These results are discussed next.

RESULTS

Table 1 shows the mean reaction time, percent correct, and eye movement measurements for each of the four visual-to-auditory transformations used in the study. Reaction times and fixation data were analyzed using a mixed model ANOVA with the Kenward-Roger correction method, and the participant's intercept as a random effect. Percent correct was analyzed using a generalized linear mixed model with a binary response distribution and a logit link function.

Note: One hundred and eighty-two trials (7.1% of all trials) were excluded from the analyses because either the participant was looking at the primary display when the alert sounded (for 11 trials) or because the participant entered an invalid combination of keys (for 171 trials). Also, each Gaze RT has been reduced by 41.5 ms to compensate for software gating time.

Table 1. Mean reaction time, percent correct, and eye movement measurements for each of the four visual-to-auditory transformations. SDs are in parentheses. Means and SDs are calculated using the sixteen participant means.

	Horizontal Transformations		Vertical Transformations	
	Square	Arc	Square	Spherical
Gaze RT (ms)	702 (335)	756 (403)	871 (457)	802 (339)
Keypress RT (ms)	1837 (233)	1789 (301)	1966 (326)	1836 (273)
Percent Correct	78.6% (6.2%)	79.9% (9.9%)	78.9% (8.1%)	68.0% (7.5%)
Number of Fixations	3.86 (0.58)	3.73 (0.64)	3.81 (0.83)	3.51 (0.62)
Gaze-to-Target Distance (in degrees of visual angle)	5.18 (0.73)	4.54 (0.88)	4.51 (0.70)	4.55 (0.65)
Number of trials	*602*	*600*	*586*	*589*

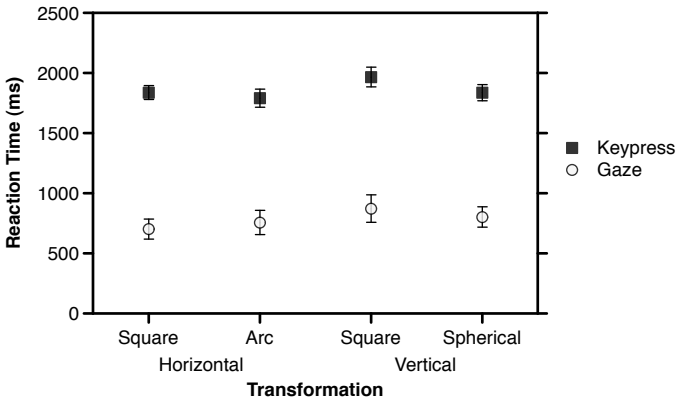

Figure 4. Reaction time for the (a) gaze arriving on the primary display and (b) keypress response. Error bars show the standard error of the 16 participant means.

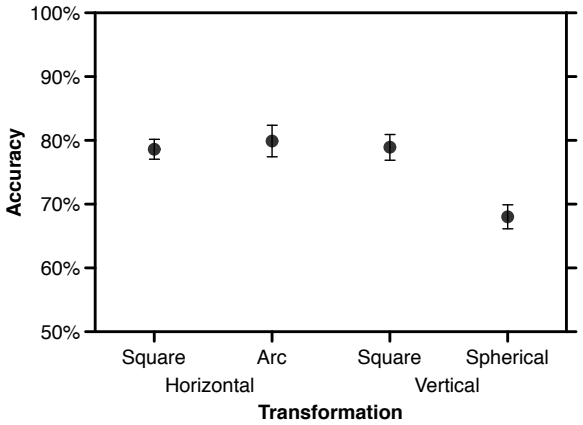

Figure 5. Accuracy as a function of transformation.

Reaction Time

For each trial, reaction time was measured from the start of the spatialized auditory alert to (a) the eyes arriving on the primary display (Gaze RT) and (b) the participant's keypress to identify the target (Keypress RT).

Figure 4 shows Gaze RT and Keypress RT as a function of each of the four visual-to-audio transformations. The transformation had a statistically significant main effect on Gaze RT, $F(3, 365)=10.18$, $p<0.0001$, with the vertical square significantly slower than all other conditions. The transformation also had a significant effect on Keypress RT, $F(3, 1496)=3.82$, $p=0.0097$. The vertical square Keypress RT was significantly slower than that of the horizontal arc and the vertical spherical surface, but not of the horizontal square. Regardless of the transformation, participants responded an average of 60 ms faster in trials in which they correctly identified the target based on the auditory alert, compared to trials in which they were incorrect, $F(1, 1897)=15.40$, $p<0.0001$.

Accuracy

The transformation had a significant main effect on accuracy, $F(3, 2350)=3.97$, $p=0.0078$. The vertical spherical surface produced significantly less accurate performance than the others, $p<0.005$, but there were no significant differences among the others. Across all transformations, accuracy was significantly affected by the distance between the target and distractor blips—more distance generated more accurate performance, $F(1, 2350)=80.70$, $p<0.0001$. The vertical position of the target on the visual display had a significant effect on accuracy—across all transformations, targets at the bottom were chosen more accurately than targets at the top, $F(1, 2350)=33.16$, $p<0.0001$.

Horizontal Versus Vertical Blip Pairs. Accuracy was analyzed specifically for fifteen target-distractor blip-pairs that were oriented either strictly-horizontally or strictly-vertically. For these trials, accuracy was significantly better with the two horizontal transformations than with the two vertical transformations, $F(1, 997)=5.26$, $p=0.0220$. The horizontal-versus-vertical orientation of these fifteen blip pairs did not have a significant effect on accuracy, $p=0.962$. However, there was a significant interaction between transformation and blip-pair orientation, $F(1,997)=11.36$, $p=0.0008$). With vertical transformations, vertically-oriented blip-pairs resulted in significantly less accurate performance than horizontal pairs, $p=0.005$. No difference in accuracy within the horizontal transformations was found, $p=0.059$.

Eye Movement Data

Number of fixations. The transformation had a significant effect on the number of fixations made in the primary display for each blip pair, $F(3, 685)=9.25$, $p<0.0001$. The vertical spherical condition yielded fewer fixations than the others, $p<0.01$ (corrected for multiple comparisons). Also, correct trials elicited an average 0.32 fewer fixations, $F(1, 2313)=29.29$, $p<0.0001$.

Gaze-to-Target Distance. The gaze-to-target distance is the distance between (a) where the gaze first lands on the primary display after the alert sounds and (b) the location of the target, which will appear milliseconds after the eyes complete that initial saccade. Since the destination of this initial saccade was influenced by the auditory but not the visual target, the gaze-to-target distance is a measure of how well participants used the auditory alert to guide their saccade to the target. There was a main effect of transformation on the gaze-to-target distance, $F(3, 422)=9.37$, $p<0.0001$. The horizontal square resulted in the greatest distance from first fixation to target, $p<0.0005$ (corrected for multiple comparisons). Correct trials elicited initial fixation points 0.55° of visual angle closer to the target than incorrect trials, $F(1, 2300)=19.40$, $p<0.0001$.

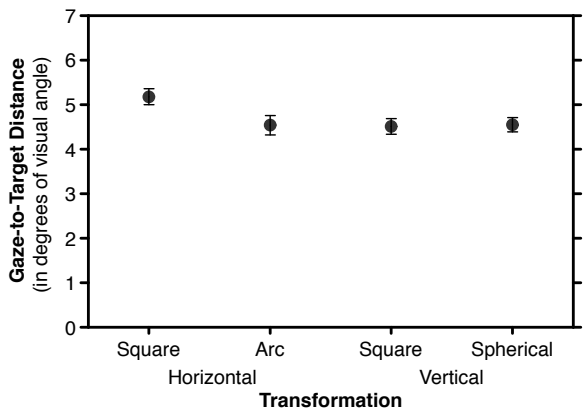

Figure 6. The gaze-to-target distance across transformations.

DISCUSSION

Across all measures, overall performance demonstrated both the potential and limitation of using the physical location of an auditory alert to convey information. Two seconds is a fairly long time to select between two visual stimuli. Accuracy of 80% is not terribly impressive when chance performance is 50%. A saccade to a location that is 4.5° from the target on a display that is just 13° by 16° of visual angle is far from an eye movement directly to the target, and will require another eye movement to fixate the target. Nonetheless, participants clearly and effectively used the spatialized audio to do the task. Further, the results when taken as a whole clearly indicate that the two horizontal transformations—especially the horizontal arc—produced better performance than the two vertical transformations.

Keypress RT and accuracy are perhaps the two most direct and important measures to consider. The vertical square Keypress RT was significantly slower than that of the horizontal arc and vertical square. The vertical spherical surface was less accurate than all of the other transformations. The poor performance with the vertical transformations cannot be explained away by a speed-accuracy tradeoff. The worse accuracy for the vertical spherical surface, for example, did not result from a more daring strategy in which participants sacrificed accuracy for speed, but resulted simply because it was harder to discern the location.

Other results help to demonstrate that the horizontal transformations outperformed the vertical. The Gaze RT for the vertical square is significantly slower than all other transformations, which suggests that more time was needed to identify the auditory location. Perhaps the strongest evidence is shown when studying the strictly-horizontal versus strictly-vertical blip pairs, in which the horizontal transformations are shown to produce more accurate performance than the vertical transformations. The interaction that appears in this analysis is also interesting. It demonstrates that the specific problem with the vertical transformations is with the vertical pairs. This is interesting because people generally have a difficult time judging the vertical position of sounds. The use of a non-individualized HRTF will exacerbate this problem, but even with an individualized HRTF, elevation cues are difficult.

The gaze-to-target distance suggests that participants had the most difficult time mapping the auditory location onto the visual display when presented with the horizontal square transformation. A rational cognitive task analysis might dictate a strategy in which participants tried to move their eyes directly to where they expected the target blip to appear. If this was the case, it was harder to do so with the horizontal square transformation.

It is interesting to consider the timeline of eye movements and keypresses as a function of the timing and duration of the alert sound. Recall that the alert sound consisted of three half-second bursts. Participants tended to move their eyes not in immediate response to the onset of the sound, but instead waited until the middle of the second burst. They then keyed in their response very shortly (a few hundred milliseconds) after the end of the third and final burst. It appears as if participants may have spent some time figuring out where the sound was before moving their eyes, or perhaps they just dutifully stuck to the joystick task until the last possible moment. It would be interesting to see how performance changes with different durations and timings of auditory alerts.

CONCLUSION

Overall, the study demonstrates that the most direct and stimulus-response-compatible mapping from a vertical visual display to a vertical auditory display will not provide the clearest auditory cues to locations on the visual display. Instead, better performance will be achieved with an auditory display that maps the horizontal coordinate on the visual display directly to the azimuth of the auditory location, and the vertical coordinate of the visual display to the perceived distance of the auditory cue.

ACKNOWLEDGEMENTS

This research is funded by the Office of Naval Research (ONR). The opinions do not necessarily reflect those of ONR.

REFERENCES

Begault, D. R. (1994). *3-D Sound for Virtual Reality and Multimedia.* Boston: AP Professional.

Cabrera, D., Ferguson, S., & Laing, G. (2005). Development of auditory alerts for air traffic control consoles. *119th Audio Engineering Society Convention,* New York, USA.

Gaver, W. W. (1997). Auditory Interfaces. In M. Helander, T. K. Landauer & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction* (2nd ed., pp. 1003-1041). Amsterdam: Elsevier.

Grantham, D. W., Hornsby, B. W. Y., & Erpenbeck, E. A. (2003). Auditory Spatial Resolution in Horizontal, Vertical, and Diagonal planes. *Journal of the Acoustical Society of America,* 114(2), 1009-1022.

Perrott, D. R., Sadralodabai, T., & Saberi, K. (1991). Aurally aided visual search in the central visual field: Effects of visual load and visual enhancement of the target. *Human Factors,* 33(4), 389-400.

Simpson, B. D., Brungart, D. S., Dallman, R. C., Yasky, R. J., Romigh, G. D., & Raquet, J. F. (2007). In-flight navigation using head-coupled and aircraft-coupled spatial audio cues. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* Baltimore, MD.

Smallman, H. S., John, M. S., Oonk, H. M., & Cowen, M. B. (2001). 'Symbicons': A hybrid symbology that combines the best elements of symbols and icons. *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting,* 110-114.

Vu, K.-P. L., & Strybel, T. Z. (2006). Effects of displacement magnitude and direction of auditory cues on auditory spatial facilitation of visual search. *Human Factors,* 48(3), 587-599.

Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using non-individualized head-related transfer functions. *Journal of the Acoustical Society of America,* 94, 111-123.

Wightman, F. L., & Kistler, D. J. (1989). Headphone simulation of free-field listening. II: Psychophysical validation. *Journal of the Acoustical Society of America*, 85, 868-878.